Philosophy & Technology (2019) 32:661–683
https://doi.org/10.1007/s13347-018-0330-6

RESEARCH ARTICLE

Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?

John Zerilli[1] · Alistair Knott[2] · James Maclaurin[1] · Colin Gavaghan[3]

Received: 9 May 2018 / Accepted: 28 August 2018 / Published online: 5 September 2018

Abstract

We are sceptical of concerns over the opacity of algorithmic decision tools. While transparency and explainability are certainly important desiderata in algorithmic governance, we worry that automated decision-making is being held to an unrealistically high standard, possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers. In this paper, we review evidence demonstrating that much human decision-making is fraught with transparency problems, show in what respects AI fares little worse or better and argue that at least some regulatory proposals for explainable AI could end up setting the bar higher than is necessary or indeed helpful. The demands of practical reason require the justification of action to be pitched at the level of practical reason. Decision tools that support or supplant practical reasoning should not be expected to aim higher than this. We cast this desideratum in terms of Daniel Dennett's theory of the "intentional stance" and argue that since the justification of action for human purposes takes the form of intentional stance explanation, the justification of algorithmic decisions should take the same form. In practice, this means that the sorts of explanations for algorithmic decisions that are analogous to intentional stance explanations should be preferred over ones that aim at the architectural innards of a decision tool.

Keywords Algorithmic decision-making · Transparency · Explainable AI · Intentional stance

✉ John Zerilli
john.zerilli@otago.ac.nz

[1] Department of Philosophy, University of Otago, Dunedin, New Zealand

[2] Department of Computer Science, University of Otago, Dunedin, New Zealand

[3] Faculty of Law, University of Otago, Dunedin, New Zealand

1 Introduction

The past decade has witnessed an unprecedented acceleration in both the sophistication and uptake of various machine learning systems, including systems that employ "deep learning" algorithms. From music and TV show recommendations, product advertising, and opinion polling to medical diagnostics, university admissions, job placement, and financial services, the range of the potential application of these new AI technologies is truly vast. Even police and law enforcement agencies have coopted deep learning tools in an effort to optimise accuracy and efficiency and reduce human bias. But while the roll-out continues to gather momentum, and enthusiasts have welcomed the dawn of a new era—the so-called "fourth Industrial Revolution" (Schwab 2016)—not everyone sees the latest iteration of AI and its steady uptake as an unmixed blessing. The concern is that, as governments increasingly automate part or all of their decision-making, the most vulnerable members of our societies may be at a significant disadvantage (Eubanks 2017; Crawford and Calo 2016). Must those awaiting the outcome of a health insurance claim, for instance, or defendants seeking bail or parole, simply take it on faith that the machine is reliable? What assurances can they be given that a human would not do a better job? Are such systems accurate, free from bias, and transparent in their operations?

Such worries are not misplaced. Many (and perhaps most) AI systems currently in use within the public sector have been acquired in the course of arms-length transactions between software development firms and government agencies. These purchasing decisions are most often regarded as routine operational matters not requiring ministerial approval. And they are being made at a time when the procurement capacities of government agencies are still not well understood (Danaher et al. 2017; Oswald and Grace 2016). While strong cultural and institutional prejudices against new technologies may also be playing a role in augmenting these worries,[1] clearly it would be remiss not to exercise some restraint here. For our part, we agree with Erdélyi and Goldsmith's (2018, p. 2) characterisation of the regulatory challenge as one that is "overwhelming… immensely complex and largely uncharted…."

But there is regulation and there is regulation. While caution is advisable, resistance to the trend of automation founded purely in concerns over the opacity of algorithmic decision-making may be missing the mark. The worry seems to be that because deep learning systems arrive at their decisions unaided, i.e. in a manner that is not specified in advance, it is not possible to interpret the system's internal processes except only approximately and imperfectly—and even this much is doubtful (Mittelstadt et al. 2016; Wachter et al. 2017a). This is thought to make such systems "inscrutable" or "opaque" in a way that is supposed to be unacceptable and anomalous when set against the capacities of human deciders, who can readily furnish specific and human-interpretable reasons in natural languages (Mittelstadt et al. 2016). While we do not deny that transparency and explainability are important desiderata in algorithmic governance, we worry that automated decision-making is being held to an unrealistically high standard here, possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers. In this paper, we

[1] Here, we have in mind certain professions that stand to lose out to automation, e.g. conveyancing, accountancy, and the like.

review evidence demonstrating that much human decision-making is fraught with transparency problems, show in what respects AI fares little worse or better, and argue that at least some regulatory proposals for explainable AI could end up setting the bar higher than is necessary or indeed helpful.

2 Explainable AI and the "Inspectability" of Algorithms

In common law countries, superior courts are usually taken to have an "inherent" jurisdiction to review decisions made by lower courts, tribunals, and administrative agencies for errors of fact or law affecting the exercise of their jurisdictions (Cane 2011; Aronson and Dyer 2013). Such reviews are "supervisory" in nature only, largely limited to the consideration of narrow questions of law. Thus, appeals against findings of fact and law more generally—whether or not going to the decider's jurisdictional competence, for example—have also been permitted in most of the world's developed legal systems since at least the nineteenth century (Baker 2002). Reviewability is plainly a concomitant of the rule of law. Moreover, since one cannot appeal a decision without knowing the bases upon which it has been reached, the transparency or explainability of a decision is likewise a crucial prerequisite of democratic governance.

Beyond the province of such formal review mechanisms, however, there are other contexts in which the reviewability of a decision will be important. Often knowing the reasons why a particular decision has been taken, even if only in rough outline, can engender trust in the process that led to it and confidence that the people in charge of the process acted fairly and reasonably (Binns et al. 2018). Transparency can thus be more than merely instrumental—the means to overturn an adverse determination, for example—and come to embody an end or democratic value in its own right, a "right to know" (Forssbæck and Oxelheim 2014; Lombrozo 2011; Heald 2006; Prat 2006). In light of such considerations, it is hardly surprising that as algorithmic decision-making technology has become more widespread, AI scholars have been increasingly concerned with the scope for reviewing algorithmic decisions. A pressing concern revolves around how best to present an algorithmic system's ratiocinations so as to be interpretable to the human subjects that may require them. The field of research concerned with this problem is known as "explainable AI" (Miller 2017), and it has become an increasingly active area of research (Pasquale 2014; Edwards and Veale 2017).

Traditional algorithms did not have a transparency problem—at least not the same one that current deep learning networks pose. This is because traditional algorithms had their rules and weights prespecified "by hand" (Mittelstadt et al. 2016), and there was nothing the system could do that was not already factored into the developer's design for how the system should operate given certain input conditions.[2] Deep learning, which is a special type of machine learning, is in a league of its own. The neural networks that implement deep learning algorithms mimic the brain's own style of computation and learning: they take the form of large arrays of simple neuron-like units, densely interconnected by a very large number of plastic synapse-like links.

[2] Traditional algorithms, like expert systems, could be inscrutable after the fact: even simple rules can generate complex and inscrutable emergent properties. But these effects were not baked in. We are grateful to an anonymous reviewer for pointing this out to us.

During training, a deep learning system adjusts the weights of these links so as to improve its performance. If trained on a decision task, it essentially derives its own method of decision-making, much as we would expect of an intelligent system. But there is the rub. In neural networks, these processes run independently of human control, so that transparency inevitably becomes an issue: it is simply not known in advance what rules will be used to handle unforeseen information. Importantly, neither the operator nor the developer will be any the wiser in this respect. *Ex ante* predictions and *ex post* assessments of the system's operations alike will be difficult to formulate precisely. This is the crux of the complaint about the lack of transparency in today's algorithms. If we cannot ascertain exactly why a machine decides the way it does, upon what bases can its decisions be reviewed? Judges, administrators, and departments of state can all supply reasons for their determinations. What sorts of "reasons" can we expect from an intelligent machine? Deep learning involves multiple hidden layers of processing that are fiendishly intricate and virtually impossible to unsnarl (Burrell 2016). Even certain simple algorithms which instantiate in the order of hundreds of rules "are very hard to inspect visually, especially when their predictions are combined probabilistically in complex ways" (Van Otterlo 2013). Added to this is the institutional context of algorithmic development. Most commercial algorithms are designed in laboratory settings consisting of numerous engineers and developers, sometimes over significant tracts of time in which personnel turnover becomes increasingly likely. In such circumstances, "a holistic understanding of the development process and its embedded values, biases and interdependencies" may simply be too much to ask (Mittelstadt et al. 2016, pp. 6–7). The problem assumes a certain immediacy the moment it is appreciated just how ubiquitous deep learning has become. Big data analytics have been used to recognise, detect, or predict speech, gestures, faces, objects, sexuality, politics, criminality, pathology, solvency, and much more. On the face of things, it is easy to sympathise with the call for greater transparency in algorithmic decision-making.

We do not deny that the new generation of deep networks introduce a distinct tension between capability and transparency in AI systems. More traditional machine learning tools, using regression methods, decision trees, or fuzzy rules, make decisions using processes that are at least somewhat inspectable. But these systems tend to be outperformed by deep networks (Muehlhauser 2013; Nusser 2009). We are also inclined to join the chorus of those demanding greater transparency in algorithmic systems. What we doubt, however, is whether the sort of transparency that lurks behind many of these demands is always useful. If human decision-making represents the gold standard for transparency, we think AI can in some respects already be said to meet it. To avoid misunderstanding, our concern is not with whether there should be some sort of "right to explanation" of algorithmic decisions, enshrined at a national or supranational level (we think there should be). Our concern is rather with the form that such rights should take. What sorts of explanations can we expect from automated decision systems, and will such explanations be good enough?

A concept that often turns up in the explainable AI literature, including in professional standards, best practice guidelines, and committee reports, is that of "inspectability" (e.g. Muehlhauser 2013; Van Otterlo 2013; Corbett-Davies et al. 2017; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2017, hereafter "the IEEE"; UK House of Commons Science and Technology Committee

2017). For instance, the IEEE's draft recommendations for algorithmic decision-making deploy this concept at numerous junctures and state that "The logic and rules embedded in the system must be available to overseers of systems, if possible. If, however, the system's logic or algorithm cannot be made *available for inspection*, then alternative ways must be available to uphold the values of transparency" (IEEE 2017, pp. 152–153, emphasis added; see also IEEE 2017, pp. 45, 71 and 180). A cognate notion is that of inspecting the "innards" or "internals" of an algorithmic decision tool (e.g. Burrell 2016; Edwards and Veale 2017, 2018; Veale and Edwards 2018; Montavon et al. 2017; IEEE 2017), also referred to as "decomposition" of the algorithm, which involves opening the black box to "understand how the *structures within*, such as the weights, neurons, decision trees and architecture, can be used to shed light on the patterns that they encode. This requires *access* to the bulk of the model structure itself" (Edwards and Veale 2017, p. 64, emphasis added; see also Wachter et al. 2017b, pp. 78–79). The IEEE (2017, p. 71) raise the possibility of designing explainable AI systems "that can provide justifying reasons or other reliable 'explanatory' data *illuminating the cognitive processes* leading to…their conclusions" (emphasis added). Elsewhere they speak of "internal" processes needing to be "traceable."[3]

It is not merely aspirational material that is drafted in this vein. Article 22(1) of the European Union's General Data Protection Regulation ("the GDPR"),[4] effective from 25 May 2018 and binding on all member states, confers upon individuals the right not to be subject to certain kinds of fully automated decisions. Article 22(3) then stipulates a range of safeguards which data controllers must implement in the event data subjects consent to such decision-making (Article 22(2)(c)). One of these, as fleshed out in a nonbinding recital (Recital 71), gives the data subject a right "to obtain an explanation of the decision reached." The term "explanation" is arguably sufficiently ambiguous to encompass types of explanation aiming at the innards of a decision tool. For instance, the Article 29 Data Protection Working Party's draft guidance on the GDPR states that "a complex mathematical explanation about how algorithms or machine-learning work," though not generally relevant, "should also be provided if this is necessary to allow experts to further verify how the decision-making process works."[5]

It is instructive to compare the kinds of explanation envisaged for predictive systems with those routinely provided by human agents. These do not yield the entrails of a decision, or "illuminat[e] the cognitive processes leading to…[a] conclusion," as the IEEE would have it (2017, p. 71). It is true that human agents are able to furnish reasons for their decisions, but this is not the same as illuminating the cognitive processes leading to a conclusion. The cognitive processes underlying human choices, especially in areas in which a crucial element of intuition, personal impression, and unarticulated hunches are driving much of the deliberation, are in fact far from transparent. Arenas of decision-making requiring, for example, assessment of the likelihood of recidivism, or the ability to repay a loan, more often than not involve significant reliance on subdoxastic factors, i.e.

[3] See, e.g. <https://standards.ieee.org/develop/project/7001.html>.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 27.3.2016, p. 1.

[5] Strictly speaking, this "good practice" recommendation (Annex 1) pertains to Article 15, not Article 22, of the GDPR. Article 15(1)(h) requires the disclosure of "meaningful information about the logic involved" in certain kinds of fully automated decisions.

factors beneath the level of conscious belief. As one researcher explains: "A large part of human decision making is based on the first few seconds and how much [the decision-makers] like the applicant. A well-dressed, well-groomed young individual has more chance than an unshaven, dishevelled bloke of obtaining a loan from a human credit checker." (Dutta 2017). A large part of human-level opacity stems from the fact that human agents are also frequently *mistaken* about their real (internal) motivations and processing logic, a fact that is often obscured by the ability of human decision-makers to invent post hoc rationalisations. Often, scholars of explainable AI treat human decision-making as epistemically privileged. For instance, Mittelstadt et al. (2016, p. 7) write that "algorithmic processing contrasts with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner's capacity to understand it." Earlier, we noted that some learning systems may be so complex that their manipulations defy systematic comprehension, and that this is most apparent in the case of deep learning systems. But the human brain, too, is largely a black box. As Muehlhauser (2013) observes:

> We can observe its inputs (light, sound, etc.), its outputs (behavior), and some of its transfer characteristics (swinging a bat at someone's eyes often results in ducking or blocking behavior), but we don't know very much about how the brain works. We've begun to develop an algorithmic understanding of some of its functions (especially vision), but only barely.

There can be little doubt that well-constructed, comprehensive, and thoughtful human reasons are extremely useful and generally sufficient for most decision-making purposes. But utility and truth are not the same things.[6] An explanation can be adequate in view of its intended audience, and yet totally inadequate from the point of view of others. Human reasons for decisions are pitched at the level of what philosophers call "practical reason"—the domain of reason which concerns the justification of action (as distinguished from "epistemic" or "theoretical reason," which concerns the justification of belief). Excessively detailed, lengthy, and technical reasons are usually not warranted, or even helpful, for most practical reasoning. Consider decisions made in the course of ordinary life. These are frequently made on the approach of significant milestones, such as attaining the age of majority, entering into a relationship, or starting a family, but they most often involve humdrum matters. Thus, whether to go out for dinner or eat in, whether to buy a new or used car, what career to pursue, whether to marry, whether and when to have children, and what to pay for a costly asset (e.g. a home, a college education), all represent typical decision points reached in a typical human life, at least in the West. Many of these decisions will be of the utmost importance to the person making them and may involve a protracted period of deliberation. But the rationales that may be expressed for them later on, perhaps after months of research or even soul-searching, will not likely assume the form of more than a few sentences. Probably there will be one factor among three or four that reveals itself after careful reflection to be the most decisive, and the stated *ex post* reasons for the decision will amount to a statement

[6] The merits of various pragmatic theories of truth are not especially relevant to us here. Another way we could put our point is that utility in the service of one aim is not utility in the service of another.

identifying that particular factor together with a few lines in defence of its putative salience.

The bulk of administrative decision-making is, we would suggest, *formally* equivalent in these respects: it will differ primarily in its content. It may concern whether to purchase new plant, whether to authorise fluoridation of the town water supply, whether to reinstate someone unfairly dismissed from a place of employment, whether to grant bail or parole, or whatever, but the decision structure is not materially different from that pertaining to ordinary life decisions. True, the stakes may be higher or lower, depending on what the decision relates to and how many people will be affected by it, and the requirement to furnish reasons—as well as the duty to consider certain factors—may be mandated in the one case and not the other. But the primary difference between administrative and ordinary life decision-making is not at the level of form. Both contexts involve practical reasoning of a more or less systematic character. And furnishing explanations that are more detailed, lengthy or technical than necessary is likely to be detrimental to the aims of transparency, regardless of the public or private nature of the situation.

It might be contended that the disanalogies between personal and official decision-making are too great to sustain this point. There are, after all, some real differences between public and private decision-making. For example, certain types of reasons are acceptable in personal but not public decision-making. It may be fine to say, "I'm not moving to Auckland because I don't like Auckland," but the same sort of reasoning would be prohibited in a public context. Furthermore, public decision-making often takes place in groups to mitigate the "noisiness" of individual reasoners, such as committees, juries, and appellate courts.[7] But these differences do not detract from their fundamentally identical structure or *form* (this is what we mean by *formal* equivalence). For either way, whether there are more or less people involved in the decision-making process (such as jurors, focus groups), or whether there are rights of appeal, both decision procedures employ practical reasoning, and take beliefs and desires as their inputs. Take judicial decision-making—perhaps the most procedurally, evidentially, statutorily, and precedentially constrained form of official reasoning that exists. Judicial reasoning is, in the first instance, supposed to appeal to ordinary litigants seeking the vindication of their rights, or, in the event of a loss, an explanation for why such vindication will not be forthcoming. So it simply must adopt the template of practical reason, as it must address citizens in one capacity or another (e.g. as members of families, as corporate executives, as shareholders, as consumers, as criminals). Even in addressing itself to lawyers, i.e. when articulating legal rules and the moral principles underpinning them, it cannot escape or transcend the bounds of practical (and moral) reasoning (Dworkin 1977, 1986).

We are not claiming that these insights are in any sense original, but we do think they are important. The decision tools coopted in predictive analytics have been pressed into the service of practical reasoning. The aim of the GDPR, for instance, is to protect "natural persons" with regard to the processing of "personal" data (Article 1). Articles 15 and 22 concern a data subject's "right" not to be subject to a "decision" based solely on automated processing, including "profiling", and the tools which have attained notoriety for their problematic biases, such as PredPol (for hot-spot policing) and COMPAS

[7] Actually, many private, purely personal decisions (regarding, e.g. what to study, which career to pursue, whether to rent or purchase) are also frequently made in consultation with friends, family, mentors, career advisers, and so on.

(predicting the likelihood of recidivism), likewise involve software intended to substitute or supplement practical human decision-making (for instance, by answering questions of the form: How should we distribute police officers over a locality having these geographical characteristics? What is the likelihood that this prisoner will recidivate?). Explanations sought from such technologies should aim for levels that are apposite to practical reasoning. Explanations that would be too detailed, lengthy, or technical to satisfy the requirements of practical reasoning should not be seen as in any way ideal.

It is somewhat remarkable then that many proposals for explainable AI assume (either explicitly or implicitly) that the innards of an information processing system constitute an acceptable and even ideal level at which to realise the aims of transparency. The UK House of Lords Select Committee on Artificial Intelligence's (2018) report is a case in point. On the one hand, what they refer to as "full technical transparency" is conceded to be "difficult, and possibly even impossible, for certain kinds of AI systems in use today, and would in any case not be appropriate or helpful in many cases" (2018, p.38). On the other hand, something like full technical transparency is "imperative" in certain safety-critical domains, such as in the legal, medical, and financial sectors of the economy. Here, regulators "must have the power to mandate the use of more transparent forms of AI, even at the potential expense of power and accuracy" (2018, p.38). The reasoning is presumably that whatever may be lost in terms of accuracy will be offset by the use of simpler systems whose innards can at least be properly inspected. Transparency of an exceptionally high standard is therefore being urged in domains where, presently, human deciders themselves are incapable of providing it. The effect is to perpetuate a double standard in which machine tools must be transparent to a degree that is in some cases unattainable, in order to be considered transparent at all, while human decision-making can get by with reasons satisfying the comparatively undemanding standards of practical reason. If simpler and more readily transparent systems are available—systems whose innards are more straightforwardly open to investigation—these should be adopted even if they produce decisions of inferior quality. And so the double standard threatens to prevent deep learning and other potentially novel AI techniques from being implemented in just those domains which could be revolutionised by them and have the most to gain. As the Committee notes (with our emphasis):

> We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate *a full and satisfactory explanation for the decisions it will take*. In cases such as deep neural networks, where it is not yet possible to generate *thorough* explanations for the decisions that are made, this may mean delaying their deployment for particular uses until alternative solutions are found. (2018, p. 40)

At the same time, and as the Committee itself noted, restricting our use of AI only to what we can fully understand limits what can be done with it (2018, p. 37).

3 Practical Reason and Dennett's Intentional Stance

We think Daniel Dennett's "intentional stance" strategy provides a useful way of clarifying the issues we have been discussing. Dennett (1987) describes three levels

of abstraction from which the behaviour of an object can be explained: the physical level, the design level, and the intentional level. These three levels are in their turn approached by adopting one of three corresponding "stances," viz.: the physical stance; the design stance; and the intentional stance. In adopting the *physical* stance towards behaviour, one provides descriptions couched in terms of the fundamental sciences, namely physics and chemistry. At this level, we are interested in features of the structure and physical constitution of the object exhibiting the behaviour in question, so that the kinds and vocabularies of these sciences will feature prominently in explanations, such as mass, velocity, and molecular arrangement. Understanding an object at this level enables us to predict its behaviour considered purely as physical stuff—a lump of matter with various structural properties—as opposed to an entity exhibiting design or more or less complex internal states (such as those possessed by persons, animals and computers). In adopting the *design* stance, by contrast, we turn from the consideration of a system's physical constitution and direct our attention to somewhat more abstract mechanical features of the system, such as its biological properties (in the case of organic systems) and engineering principles (in the case of built artefacts). At this level of inquiry, we are concerned with how the object functions as an integrated mechanism, i.e. how its parts cohere to generate systematic behaviour of a specific sort. The study of anatomy requires the adoption of a design stance towards the behaviour of the human body, as does the study of low-level programming languages towards the behaviour of a modern computer. As Dennett puts it:

> If you know something about the design of an artefact, you can predict its behavior without worrying yourself about the underlying physics of its parts. Even small children can readily learn to manipulate such complicated objects as VCRs without having a clue how they work; they know just what will happen when they press a sequence of buttons, because they know what is designed to happen. They are operating from what I call the design stance. (Dennett 1995, p. 229)

Lastly, and most abstractly, in adopting the *intentional* stance, we eschew all considerations of physical structure—of ions and molecules and valency—and all considerations of biology and engineering—of the functional properties of the various components of a mechanism inhering in its design—and describe the system purely in terms of mental states, i.e. folk concepts, propositional attitudes, and belief-desire psychology. This is the stance from which we understand ordinary human behaviour and engage in practical reasoning. For example, if we wish to explain why Mary decided to stay home rather than go out on Saturday night, we would typically do so by adopting the intentional stance, employing the kinds and vocabulary distinctive of the intentional level. We would say that Mary decided to stay in because she had an expensive week the previous week and does not want to break her budget because she is trying to save money to buy a home. This is a very powerful explanation, despite the fact that it makes no reference to biological, engineering, or microstructural details and is in fact composed entirely of beliefs and desires (e.g. the belief that too much money was spent last week, that one must save in order to buy a home, the desire to own a home). Indeed if we sought to explain Mary's behaviour using the resources available from the standpoint of the design or physical levels, we would be either at a complete loss, or—with the appropriate expertise to hand—bogged down in an extremely complex and vast

array of descriptions formally cataloguing her brain states at every point in her decision-making. Complex descriptions descending to such fine details as electrochemical transduction patterns and the state of billions of neurons would clearly be inappropriate for practical reasoning. Neither others seeking to understand Mary's behaviour on Saturday night nor she herself in accounting for her decision to stay at home would think to provide explanations otherwise than at the intentional level. As Dennett (1987, p. 17) states: "A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do." And what we can say of Mary here applies no less to an administrative decision-maker or judicial officer. There is once again, we maintain, formal equivalence in the two types of decision-making. Only the content will significantly differ in the two cases. In the case of a judge deciding on the appropriate sentence for the perpetrator of sexual assault, for instance, the beliefs will obviously be more numerous and more varied than in Mary's decision—beliefs as to what the proper interpretation of the law governing sexual assault should be, beliefs as to the relevance of being a victim of sexual abuse in the past on the propensity to engage in sexually violent acts in the future, and so on.

A number of features of intentional stance explanations are worth calling attention to. Intentional stance psychology is essentially commonsense or folk psychology, and a good deal of work in the philosophy of language bears directly on the desiderata for explanations in just this domain. Such explanations

- tend to be *contrastive*—people tend not to inquire why event P occurred, but rather why P occurred *instead of Q*
- tend to be *selective*—people do not expect an exhaustive reckoning of all factors and causes, only the significant ones
- are invariably social—they involve a type of interaction between at least one person (an explainer) and at least one other person (explainee) (Miller 2017, pp. 5–6)

These features highlight the important fact that folk psychological explanations, because they concern interpersonal matters, often fasten upon *reasons*, as opposed to *causes*. Unconscious motives, emotions, culture, personality, and surrounding context are causes that may *lead* to reasons (Miller 2017). But whereas reasons "belong to the actor," causes themselves do not (Miller 2017, p. 25). Our language reflects this distinction. We tend not to ask "What were Mary's *causes* for staying in?" but rather "What were Mary's *reasons* for staying in?" We could well ask, "What *caused* Mary to stay in?" or "What *causes* led to Mary staying in?" But such questions do not straightforwardly direct attention to her motivations as a rational agent—unconscious, impersonal causes do not *belong* to Mary in the same way her reasons do, if they do at all. Additionally, folk psychological explanations are primarily *conversational*, which is what distinguishes them from merely causal explanations (Hilton 1990). This means that to provide a folk psychological explanation is literally to engage in a kind of conversation. In turn, such conversations will tend to follow certain settled conventions, best captured in Paul Grice's (Grice 1975) well-known maxims. The most pertinent for our purposes are (roughly): say only what you believe, say only what is necessary, and say only what is relevant. According to Hilton (1990), folk psychological explanations should strive to adhere to these maxims. To provide "too much information" in your explanations—information that is not necessary, or information that is simply not

relevant for conveying the essential reasons for your decision—is to violate the Gricean maxims. Design level and physical level explanations run a real risk of violating the maxims, for such explanations provide information that is hardly necessary or relevant for those seeking reasons (as opposed to causes).

The ontological status of intentional stance/folk psychological items such as beliefs and desires is a matter of intense controversy in the philosophy of mind (Churchland 1981; Fodor 1981; Dennett 1991; Stich 1983). Are such items "real" in the sense of having the same ontological and epistemic purchase of items at the physical and design levels? We once thought that the mentally ill were possessed by evil spirits. As the sciences of psychology and neuroscience continue to mature, will we one day come to think it preposterous to regard beliefs and desires as any more real than evil spirits? In view of how much we rely on folk psychology for day-to-day living, it is a unwise to be smug about these issues. Fortunately, we do not have to take a position on the matter to appreciate the value of Dennett's intentional strategy. At a minimum, intentional stance categories are extremely useful, adhering to Eleanor Rosch's (1978, p. 28) condition of providing "maximum information with the least cognitive effort," as well as the Gricean maxims. Hence, we propose that it is best to think of intentional stance categories—and the human reasons they enter into—as providing useful schemata of the complex neurophysiological phenomena to which they relate. The difficulty (and indeed present *impossibility*) of obtaining physical level explanations for human decisions does not prevent our making full use of intentional level explanations. And even though design level explanations of human decision-making are available (see sect. 5), these have not yet advanced to the point where they are able to assist in meeting the real-time demands of human decision-making. Human decision-makers, at least to our knowledge, have never been required to furnish anything like design level explanations for their decisions.

Crucially, these same considerations apply to deep learning and other "opaque" decision tools (Chopra and White 2011, Ch. 1). The intentional stance can be adopted towards the behaviour of any system with internal representational states whose transitions can be described using formal rules. Turing machines and von Neumann devices whose inner states transition in response to inputs manipulated in accordance with formal rules are clear candidates for intentional stance explanations. In the case of computers, however, the items of the intentional stance will normally consist of the syntactic elements of high-level programming languages rather than beliefs and desires per se. As we discuss in Section 6, serviceable schemata cast at the intentional level are already available for deep learning systems and continue to develop rapidly. These ought to suffice for purposes of algorithmic transparency, even though they probably fall well short of true technical transparency at the design level.

4 Bias and Transparency in Algorithmic Decision-Making

Research into human decision-making has generated many important results over the past thirty years (Pomerol and Adam 2008), but one of the most critical findings from our point of view concerns the central role that human emotions play in decision-making (Damasio 1994). The centrality of emotions in human decision-making at once suggests a unique contribution automated decision technology can make to practical

reasoning: it can significantly reduce one of two potential sources of bias and discrimination. This is bias that resides in a system by virtue of its design, structure, and rules of operation, or as a consequence of inputs effecting a permanent change in its design, structure, and rules of operation. We term this *intrinsic* bias.[8]

Human bias is often intrinsic, in the above sense, because it bears an important relation to emotion, itself a constitutive feature of personality (Angie et al. 2011; Pohl 2008; Stephan and Finlay 1999). Racial bias is a good example of intrinsic bias in human beings, because the connection with emotion is relatively clear (the emotion being fear), as is its tolerance to falsifying evidence. When someone has been conditioned to believe that an ethnic minority poses a threat to safety, or is more susceptible to crime, merely supplying that person with evidence to the contrary may be insufficient to dislodge a lifetime of encrusted prejudice (Bezrukova et al. 2016). Racist conditioning may permanently (or semipermanently) affect the way a person processes information and makes decisions. Of course this is not to say that intrinsic bias is always irrational. Many human biases could be thought to result from the misfiring of an ancient and conserved cognitive adaptation to make generic judgements (Begby 2013; Leslie 2017). Because such judgements are based on dispositional rather than probabilistic factors, they too tend to be resistant to disconfirming evidence.

As against intrinsic bias, bias that is *not* intrinsic (i.e. extrinsic) derives from a system's inputs when they do not effect a permanent change in the system's internal structure and rules of operation. In these cases, false information may affect a system's outputs, but so long as the information is corrected, the outputs will be unbiased pro tanto. Thus, if a person is given information that leads them to the erroneous belief that p, and the belief that p plays a relevant role in decision-making, leading to the decision that q, the person will be nonintrinsically biased towards the decision that q if, upon receiving the correct information, the person no longer believes that p, and either abandons or revises the grounds for the decision that q.

Overall, while it is true that an algorithm can be intrinsically biased (see below), nonintrinsic bias is probably the bigger issue for AI (Friedman and Nissenbaum 1996; Johnson 2006). The so-called dirty data problem is a neat illustration. Errors and biases latent in data training sets tend to be reproduced in the outputs of machine learning tools (Barocas and Selbst 2015; Diakopoulos 2015). This is a significant problem, and one that is compounded—of all things—by copyright and intellectual property laws, which presently limit the access users have to better quality training data (Levendowski 2017).[9] But nonintrinsic bias is still in principle less difficult to overcome than intrinsic bias. Most of these problems arise from the use of unrepresentative data sets. For instance, face recognition systems trained predominantly on Caucasian faces might reject the passport application photos of Asian persons, whose eyes appear closed

[8] Our citing Damasio (1994) might seem odd, for we are suggesting that the effects of emotions may be reason-distorting, whereas for Damasio this is not the main point. Damasio sees emotions as an *essential* component of rational thought (and we agree). Nevertheless, he does see emotions as engendering biases in some cases. For instance, he says: "I will not deny that uncontrolled or misdirected emotion can be a *major* source of irrational behavior. Nor will I deny that seemingly normal reason can be disturbed by subtle biases rooted in emotion" (1994, pp. 52–53, our emphasis). (He goes on to say: "Nonetheless, (…) [r]eduction in emotion may constitute an equally important source of irrational behavior." But the key point is that he does see emotions as a potential source of bias in some contexts.)

[9] Copyright law is not the only culprit here. Other factors impeding access include privacy and income disparities.

(Griffiths 2016). Speech recognition systems, too, are notorious for being less accurate when decoding female voices than male ones (Tatman 2016). Both situations arise from a failure to include members of diverse social groups in training data. The obvious solution is to diversify the training sets (Klingele 2016; Crawford and Calo 2016). While there are political and legal barriers in the way of this, as Levendowski (2017) documents in her careful analysis of intellectual property laws, it is not nearly as intractable a problem as the one posed by intrinsic human bias (Bezrukova et al. 2016; Plous 2003a; Allport 1954).

Of course not all dirty data suffers from being unrepresentative. For instance, a machine learning tool that disproportionately classifies African Americans as posing a greater risk of recidivism has probably learnt from a data set that reflects racial prejudices inherent in previous discriminatory patterns of policing (Larson et al. 2016; Lum and Isaac 2016; Crawford and Calo 2016). Again, this would not count as intrinsic bias, on our definition, because the data do not affect the system's internal structure and rules of operation. But nor can such bias be said to originate from "unrepresentative" data, which can in theory be corrected by including more diverse ethnic groups in the training set. The bias here stems from intrinsic *human* bias, with machines simply inheriting the bias from prevailing social conditions. There is another side to the story too. It seems that fairer algorithms are not possible that satisfy any more than one definition of fairness at a time, because "many notions of fairness are in conflict" (Corbett-Davies et al. 2017, p. 799; see also Hardt et al. 2016; Kleinberg et al. 2017). Complicating the matter further, public safety and fairness also collide. As Corbett-Davies et al. (2017) conclude after a rigorous statistical examination of the issue:

> satisfying common definitions of fairness means one must in theory sacrifice some degree of public safety….Maximizing public safety requires detaining all individuals deemed sufficiently likely to commit a violent crime, regardless of race. However, to satisfy common metrics of fairness, one must set multiple, race-specific thresholds. There is thus an inherent tension between minimizing expected violent crime and satisfying common notions of fairness. (2017, pp. 802, 804)

Importantly, they note that "the principles we discuss apply to other domains, *and also to human decision makers carrying out structured decision rules*" (2017, p. 797, emphasis added); similarly, "there is a mathematical limit to how fair any algorithm—*or human decision-maker*—can ever be" (Corbett-Davies et al. 2016, emphasis added). Given that this feature of nonintrinsic bias bedevils every decision system, we see it as providing no justification for applying higher standards of transparency to algorithmic decision systems.

What about intrinsic algorithmic bias? We mentioned in passing that algorithms can be *intriniscally* biased too (like humans). Our reasoning is that algorithmic development is never an entirely objective, value-free endeavour: it will be influenced by a host of social and institutional norms, practices and attitudes that could well build bias into design. The social and institutional factors we have in mind here include—but are not limited to—the predominantly white, technically educated, and male composition of the field of AI (Crawford 2016). At the same time, we suggested that intrinsic bias poses less of a problem for AI than it does for humans. In light of the insidious effects

that social and institutional factors may play in shaping the design of algorithms, our suggestion could seem misguided. For it to be true, the influence of social and institutional factors on algorithmic development would have to be less pronounced than the effects of social conditioning on human life (e.g. the effects of socialisation on the development of racist attitudes). This is certainly what we have assumed. Given that algorithmic decision tools are not persons, but rather built artefacts with a far less complex internal structure than human beings, no independent self-sustaining culture or framework of embedded values, and no emotional capacity at all, we think that the assumption is a fair one—provided we exclude from consideration tools which have been *consciously* designed to inflict harms on groups of people (e.g. lethal autonomous weapon systems).

Even so it may be countered that some types of intrinsic bias are simply par for the course in algorithmic systems. For example, what Friedman and Nissenbaum (1996) call "technical" bias arises from the inherent constraints imposed by the technology itself. Mittelstadt et al. (2016, p. 7) give as examples "when an alphabetical listing of airline companies leads to increase [*sic*.] business for those earlier in the alphabet, or an error in the design of a random number generator…causes particular numbers to be favoured." Intrinsic biases may also emerge from advances in knowledge, in the way medical diagnostic tools that do not account for new knowledge will be "unavoidably biased towards treatments included in their decision architecture" (Mittelstadt et al. 2016, p. 8). While we do not deny these forms of bias, again we are not convinced that they are unique to AI. In fact there is bound to be intrinsic bias of this technical and emergent sort in *any* decision system, be it natural or engineered. Tversky and Kahneman's (1974) "availability heuristic" in human decision-making is very analogous to the alphabetical listing problem Mittelstadt et al. (2016) cited to illustrate technical bias. Consider also that professionals such as medical practitioners, lawyers, and tax agents must maintain a certain standard of knowledge in order to be considered proficient and that this is generally enforced through mandatory continuing education programs. This is an open avowal of the fact that humans are not immune to emergent bias either. Hence, once again the existence of machine bias on its own cannot justify the imposition of a higher standard of transparency for AI. Standards for machines taking the form of mandated software upgrades and maintenance procedures would be analogous to mandatory continuing education programs for professionals and would probably solve the clinical diagnostics problem which Mittelstadt et al. (2016) also cited. A consistent standard of transparency across the board is therefore possible in principle and seems reasonable in the circumstances. We will return to this point in Section 7.

5 Unconscious Biases and Opacity in Human Decision-Making

Plous (2003b, p. 2) observes early in his survey of human prejudice that "humans are cognitively predisposed to harbor prejudice and stereotypes." He later goes on to observe that "contemporary forms of prejudice are often difficult to detect and may even be unknown to the prejudice holders" (Plous 2003a, p. 17). More recent research has corroborated these observations. It seems that the tendency to be unaware of one's own predilections is even present in those with regular experience of having to handle incriminating material in a sensitive and professional manner. In a recent review of

psycho-legal literature comparing judicial and juror susceptibility to prejudicial publicity, the authors note that although "an overwhelming majority of judges and jurors do their utmost to bring an impartial mind to the matters before them…even the best of efforts may nonetheless be compromised" (McEwen et al. 2018, p. 126). They write that "even accepting the possibility that judges do reason differently to jurors, the psycho-legal research suggests that this does not have a significant effect on the fact-finding role of a judge," (McEwen et al. 2018, p. 136) and that "in relation to prejudicial publicity, judges and jurors are similarly affected" (McEwen et al. 2018, p. 140). This should force us to reassess our attitudes to human reasoning, and question the capabilities of even the most esteemed reasoners. The practice of giving reasons for decisions may be simply insufficient to counter the influence of various factors, and the reasons offered for human decisions could well conceal motivations scarcely known to the decision-makers. Even when the motivations *are* known, the stated reasons for a decision can serve to cloak the true reasons. In common law systems, it is well-known that if a judge has decided upon a fair outcome, and there is no precedent to support it, the judge may well grope around until *some* justification can be extracted from what limited precedents do exist (Waldron 1990). Furthermore, rights of appeal are limited. People forget that substantial parts of judicial reasoning are essentially inviolable, even in the lowest courts. Judicial *discretion* is a quintessential black box that can often only be appealed within severely narrow limits.[10] Given how frequently judges are called upon to exercise their discretion, this could be seen as contrary to the principles of open justice. Judges are also allowed considerable leeway in respect of their findings on witness credibility. Appellate courts are generally reluctant to overturn judicial determinations of credibility, because the position of trial judges in being able to assess the demeanour of a witness at first hand is seen to deserve particular respect.[11] Finally, it is worth remembering that appeals rarely lie as of right anyway—often the rules of civil procedure will restrict the flow of appeals from lower courts by requiring the appellate court to grant leave first.[12]

Even without taking a stand on the question of free will, the purely neurophysiological aspects of human decision-making are not understood beyond general principles of interneural transmission, excitation and inhibition. In multi-criterion decision cases, where a decision-maker must juggle a number of factors and weigh the relevance of each in arriving at a final decision, one hypothesis has it that the brain eliminates potential solutions such that a dominant one ends up inhibiting the others in a sort of "winner takes all" scenario (Pomerol and Adam 2008, p. 24). While this process is to some extent measurable, "it is essentially hidden in the stage where weights or relative importance are allocated to each criterion" (Pomerol and Adam 2008, p. 24). It serves as a salutary reminder that even when a sentencing judge provides reasons allocating weights to various statutory factors, the actual inner processing logic behind the allocation remains obscure.

More general work on the cognitive psychology of human decision-making is no less sobering. "Anchoring" and "framing" effects are well-known to researchers in the field. One such effect, the "proximity" effect, results in more recent events having greater weight than older ones and bearing a greater influence on choices in the search

[10] *House v. The King* (1936) 55 C.L.R. 499 (High Court of Australia).

[11] *Devries v. Australian National Railways Commission* (1993) 177 CLR 472 (High Court of Australia); *Abalos v. Australian Postal Commission* (1990) 171 CLR 167 (High Court of Australia); cf. *Fox v. Percy* (2003) 214 C.L.R. 118 (High Court of Australia).

[12] See, e.g. Supreme Court Act, s. 101(2) (New South Wales).

for solutions (Pomerol and Adam 2008). The tendency to see false correlations where none exists is also well documented (Piattelli-Palmarini 1995; Tversky and Kahneman 1974). The bias is at its strongest when a human subject is having to deal in small probabilities (Pomerol and Adam 2008). Furthermore, constraints imposed by short-term memory capacity mean we cannot handle more than three or four relationships at a time (Pohl 2008). Because it is in the nature of complex decisions to present multiple relationships among many issues, our inability to assess these factors concurrently constitutes a significant limitation on our capacity to process complexity.

6 Explainable AI 2.0

We have suggested that because the demands of practical reason require the justification of action to be pitched at the level of practical reason, decision tools that support or supplant practical reasoning should not be expected to aim for a standard any higher than this. We cast this desideratum in terms of Dennett's intentional stance and argued that since the justification of action for human purposes takes the form of intentional stance explanation, the justification of algorithmic decisions should take the same form. In practice this means that the sorts of explanations for algorithmic decisions that are analogous to intentional stance explanations should be preferred over ones that aim at the architectural innards of a decision tool. In this section, we provide a rough sketch of what these analogues might look like.

Perhaps the most useful thing a decision subject wants to know is how different factors were weighed in coming to a final decision. It is common for human decision-makers to disclose these allocations, even if, as we mentioned earlier, the inner processing logic leading to them remain obscure. Weights are classic exemplars of intentional stance logic, and one way for algorithmic decision tools to be held accountable in a manner consistent with human decision-makers is by having them divulge their weights (Montavon et al. 2017). As Edwards and Veale (2018) remark, "Extracting estimates of the weightings within a complex algorithm is increasingly possible, particularly if only the area 'local' to the query is being considered." This is because local terrain, "unlike the complex innards of the entire network, might display recognisable patterns" (Edwards and Veale 2018). It is therefore heartening to see the development of various model-agnostic explanations that provide pedagogical guidance, or "models-of-a-model" (Edwards and Veale 2017). Rather than opening the black box, which runs the familiar risk of disclosing proprietary code, pedagogical techniques work by "querying" the system, for example, through a trace program or test routine (Chopra and White 2011). These models are directly relevant to our analysis, because intentional stance explanations are in some respects themselves "models of a model," providing "real patterns" of complicated phenomena more comprehensively described at the design level (Dennett 1991).

To make such explanations as user-friendly as possible, once local variables have been extracted, it may be possible to format them to one of a number of explanatory styles likely to be useful to an explainee/end-user. Binns et al. (2018) reviewed the literature on interpretable machine learning models, together with legal commentary on the GDPR's requirement that "meaningful information about the logic involved" be disclosed to data subjects significantly affected by decisions reached solely through automated means (Articles 15(1)(h) and 22(1)). Their aim was to distil a serviceable set of explanation styles that would comply with both the technical and legal desiderata of

explainability contained in these materials. They settled upon four distinct explanation styles: (i) *input influence*-based explanations indicate the influence of a range of factors on the outcome; (ii) *demographic*-based explanations reveal characteristics of those who were similarly classified; (iii) *case*-based explanations present the characteristics of another decision subject with the same outcome; and finally, (iv) *sensitivity*-based explanations specify factors about the decision subject which would need to change for the outcome to be different (see Box 1 for examples). Crucially, each of these explanation styles, but particularly input influence- and sensitivity-based styles, conveys information pitched at the intentional level and commensurate with the demands of practical reason. Indeed input influence-based explanations possess the rudimentary structure of judicial reasons for factual findings and even judicial remarks on sentence.

Box 1

Example explanation styles

(i) Input influence-based explanations

Our predictive model assessed your personal information and driving behaviour in order to predict your chances of having an accident. The more + s or − s, the more positively or negatively that factor impacted your predicted chance of accidents. Unimportant factors are indicated.

> Your age (−−−)

> Driving experience (−−−)

> Level of adherence to speed limit (−)

> Number of trips taken at night (++)

> Miles per month (+)

(ii) Demographic-based explanations

> 29% of female drivers qualified for the cheapest tier

> 31% of drivers in your age group [30–39] qualified for the cheapest tier

> 35% of drivers with 17 years of experience qualified for the cheapest tier

> 15% of drivers who have been in one accident which was not their fault qualified for the cheapest tier

> 26% of drivers who regularly travel at night qualified for the cheapest tier

> 21% of drivers who exceed the speed limit once over two months qualified for the cheapest tier

(iii) Case-based explanations

This decision was based on thousands of similar cases from the past. For example, a similar case to yours is a previous customer, Claire. She was 38 years old with 18 years of driving experience, drove 850 miles per month, occasionally exceeded the speed limit, and 25% of her trips took place at night. Claire was involved in one accident in the following year.

(iv) Sensitivity-based explanations

> If 10% or less of your driving took place at night, you would have qualified for the cheapest tier.

> If your average miles per month were 700 or less, you would have qualified for the cheapest tier.

Source: Binns et al. (2018, p. 6)

7 Double Standards: Good or Bad?

Auditing protocols serve the aims of "transparency" construed in a much broader sense than explainability or interpretability alone. They often extend across the full gamut of both *ex ante* and *ex post* decision contexts and are designed to promote confidence and

public trust in extant decision-making regimes. In driving home the main argument of this section, it may therefore be useful if at the start we distinguish between two quite different auditing solutions that could be adopted for decision systems satisfying different needs and posing quite different levels of risk. Thus, we distinguish between what we call *performance*-based and *accreditation*-based auditing models.[13] Performance-based auditing, as we define it, is outcome oriented: can the system perform the work properly, as a simple matter of fact? This requires periodic monitoring in the form of independent reviews, annual reports, mandatory continuing education or software upgrades (in the case of machines), and/or the publication of official reasons for decisions. This represents the Rolls Royce standard. Accreditation-based auditing, by contrast, is expertise oriented: does the system have the appropriate qualifications and pedigree to be entrusted to perform the work properly, as perhaps evidenced by a relevant tertiary qualification, or (in the case of machines) a pre-procurement certification scheme? This obviously represents a lower auditing standard. We imagine these two models situated at opposite ends of a continuum. Some contexts are especially sensitive and require a strong performance-based auditing regime to be used in conjunction with a certification scheme of some kind (e.g. social work and community services, medical screening, parole decisions). In other contexts (e.g. recommender systems, web page ranking algorithms), the weaker accreditation-based standard may be all that is required. The crucial point is that the standards of transparency, even in its widest sense, can and—without some compelling political, economic or social justification to the contrary—*should* be applied consistently across the board, regardless of whether we are dealing with machines or humans. In higher stakes decision settings, the Rolls Royce standard should apply. In lower stakes settings, an omnibus standard will do. The kind of decision regime in place makes little difference to the standard of transparency we should expect, given the stakes involved.

This leads directly on to the question: just what sorts of countervailing political, economic, or social factors *would* justify the application of different standards? One factor we can think of—with resonances of the political, economic, and social all at once—is the potential of AI to advance well beyond the level of which humans are presently capable in a particular domain. For instance, what if algorithmic decision tools have a good chance of being significantly better than humans, not just in the area of explainability, but as regards accuracy, bias, and so on, and not just by a small margin, but by a considerably wide margin? Should regulations then be crafted with a view to bringing out the best that AI can be, even if this means setting a regulatory standard that would be far stricter than would ever apply to a human being? In our view, if AI advances to the point where it will be significantly more accurate and less discriminatory than human decision-making, a double standard would probably be justified. For aside from efficiency gains (a compelling economic and political consideration in its own right), what reasons would we have for implementing algorithmic decision tools if they were no more accurate or fair than human beings in deciding the same questions? Thus, in taking the tough line we have on double standards, we do not mean to imply that there will never be contexts in which AI can be held to higher regulatory standards than human beings. More than likely there will be such contexts. But in

[13] This classification is not to be confused with the more traditional one found in the standards literature, e.g. Coglianese and Lazer (2003).

the explainable AI literature, to the best of our knowledge, no one has argued that algorithmic decision tools have a greater potential for transparency than human beings. On the contrary, the prevailing attitude towards AI on this issue has been condescending. And besides, as we have been urging, more information is not always ideal. So at least one important justification for double standards in explainable AI does not arise.

Are there others? Here are potentially three.[14] First, it might be thought that a single algorithm can have a much bigger impact than a single human decision-maker, inasmuch as a single algorithm can "rule" over millions of people, whereas a single human decision-maker's determinations typically extend no farther than the (small) number of people appealing to his or her jurisdiction. But we very much doubt this. Some offices are occupied by persons wielding unseemly influence in international affairs (it is trite to point them out), and the ramifications of their decisions extend far and wide. Moreover, jurisdictions are a little like clades in biology: a national supreme court may encompass the jurisdictions of various state or provincial supreme courts combined. Every decision it hands down will affect not only the litigants before it but also all others placed in similar legal circumstances. Indeed because its rulings are law, and final, they bind every citizen of the state. Besides, even if it were true that algorithms had a bigger impact than human decision-makers, in itself, this would not justify imposing a higher standard of transparency in principle. For having a greater impact simply means the stakes are higher. And when the stakes are higher, as we said, a Rolls Royce standard should apply, *regardless of whether a human or machine is involved*. If an office-holder is authorised to deploy the armed forces of a state, thereby affecting the lives not only of the soldiers deployed but also of their families and the country at large, surely a higher degree of transparency should be applied to them than to a person—or machine—whose decisions are not attended by such consequences. To repeat, the kind of decision regime in place (be it natural or artificial) makes little difference to the standard of transparency we should expect, given the stakes involved. Prima facie, standards of transparency ought to be sensitive to the *stakes* of a decision, not to *who* or *what* is making it.

Another justification for higher standards might make something of the fact that algorithms cannot be held either *interpersonally* or (at this stage) *legally* accountable for their mistakes. If a human errs, he or she is at risk of dismissal, disgrace, fines, or imprisonment. These consequences serve (in part) to incentivise good behaviour and depend on the responsiveness of the agent to reasons and affective attitudes. Algorithms lack this responsiveness and hence (so the argument goes) should be subjected to a higher standard of transparency than human decision-makers. Again, however, we remain unconvinced that this has any bearing on standards of transparency. The need for such incentives in the first place arises precisely from the possibility that human motivations can lead people astray. Machines do not have to be incentivised to behave properly *just because* they cannot be incentivised to behave poorly. In fact, it would not be unreasonable to suggest that counteracting perverse incentives through censure, penalties, and other sanctions is a way for human beings to be brought *up* to the standard of machines. Such measures do not give human beings a head start—rather, they eliminate the advantage which autonomous systems already have on human decision-makers.

[14] We are grateful to an anonymous reviewer for bringing these to our attention.

A final consideration might proceed as follows. The kind of decisions we are worried about when discussing algorithmic decision-making are decisions regarding policies that affect third parties. In these kinds of decisions, procedures are in place to minimise individuals' biases, such as expert reports, committees, and appeal mechanisms. And this might be thought to tip the scales in favour of human decision-making, justifying a more lenient standard of transparency.

In Section 5, we argued that appeal mechanisms are restrictive and limited in their potential to reduce bias. Regarding committees, we cited a recent paper demonstrating that both juries (a type of committee) and judges are vulnerable to prejudicial media publicity. So just having more people involved in a decision does not necessarily eliminate or reduce the potential for human bias to interfere with human reasoning. As for expertise, judges are a type of expert, and as we said, even when their own motivations are known, the stated reasons for their decisions can serve to cloak the true reasons.

But the point about committees is well taken. Here, the thought is that a high standard of transparency is naturally enforced by processes within a group, because members often need to justify and rationalise their points of view, which are typically challenged or queried in the ordinary course of discussion. Nevertheless, research in social psychology suggests that the group-based mechanisms which ensure the *production* of justifications do not always guarantee their *quality*. In fact, participants in a group are often swayed by the mere presence of a justification, regardless of its quality. In a classic study, Langer et al. (1978) found that intrusions into a photocopier queue were more likely to be tolerated if a justification was provided, even if it was devoid of content. "May I use the Xerox machine, because I have to make copies?" was more effective than "May I use the Xerox machine?" (with no "because…"). Of course the result speaks directly to the dynamics of an informal group setting, not a high-level public committee. But it has been taken seriously by legal theorists in discussions of legitimacy (see, e.g. Oliver and Batra 2015, p. 72, for the discussion). Thus, group processes, which naturally elicit justifications, do not necessarily improve on solo decision-making. And what is more, even it could be shown that a *single* machine's decisions were less transparent than those made by a group of people, this would seem less a shortcoming of algorithms than an asymmetry in the systems being compared. A decision made by *one* person would, for the same reason, be less transparent than a decision made by a *group* of people.

8 Conclusion

We have tried to expose an assumption implicit behind some of the proposals and regulations around explainable AI which seek to make algorithmic decision tools more transparent. The assumption seems to be that it is fair to impose a higher standard of transparency on such tools than would ordinarily be imposed on human decision-makers. Either that or the assumption is simply that human decisions are comparatively more transparent than algorithmic decisions, because they can be inspected at a depth to which AI is not presently amenable. We have argued that both assumptions are false. At this stage, the sorts of explanations we cannot obtain from AI we cannot obtain from humans either. Subtle biases, subdoxastic cues, and unconscious predilections may lie

well beneath the reach of introspection or simply evade explicit recognition in official reasons. Fortunately, however, the sorts of explanations we *can* expect to obtain from human beings we may be able to obtain, mutatis mutandis, from AI systems too, and these really ought to satisfy the demands of explainable AI.

Acknowledgments The authors wish to thank the participants of two roundtables, one held in Oxford, November 23–24, 2017, in partnership with the Uehiro Centre for Practical Ethics, University of Oxford, and the other in Dunedin, December 11–12, at the University of Otago.

Funding This research was supported by a New Zealand Law Foundation grant (2016/ILP/10).

Compliance with Ethical Standards

Conflict of Interest AK works for Soul Machines Ltd under contract. JZ, JM, and CG have no other disclosures or relevant affiliations apart from the appointments above.

References

Allport, G. W. (1954). *The nature of prejudice*. Cambridge: Addison-Wesley.

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: a meta-analytic review. *Cognition and Emotion, 25*(8), 1393–1422.

Aronson, & Dyer. (2013). *Judicial review of administrative action* (5th ed.). Sydney: Lawbook Co..

Baker, J. H. (2002). *An introduction to English legal history* (4th ed.). New York: Oxford University Press.

Barocas, S., & Selbst, A. D. (2015). Big data's disparate impact. *California Law Review, 104*, 671–732.

Begby, E. (2013). The epistemology of prejudice. *Thought, 2*(2), 90–99.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016) A meta-analytical integration of over 40 years of research on diversity training evaluation. Available at: http://scholarship.sha.cornell.edu/articles/974.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. & Shadbolt, N. (2018) "It's reducing a human being to a percentage": perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York: ACM.

Burrell, J. (2016). How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 1–12.

Cane, P. (2011). *Administrative law* (5th ed.). New York: Oxford University Press.

Chopra, S., & White, L. F. (2011). *A legal theory for autonomous artificial agents*. Ann Arbor: University of Michigan Press.

Churchland, P. A. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67–90.

Coglianese, C., & Lazer, D. (2003). Management-based regulation: prescribing private management to achieve public goals. *Law and Society Review, 37*(4), 691–730.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2016) Algorithmic decision making and the cost of fairness. Proceedings of KDD'17. Available at: https://arxiv.org/pdf/1701.08230.pdf.

Corbett-Davies, S., Pierson, E., Feller, A. & Goel, S. (2017) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*.

Crawford, K. (2016) Artificial intelligence's white guy problem. *New York Times*.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*, 311–313.

Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain*. New York: Putnam's Sons.

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., & Shankar, K. (2017). Algorithmic governance: developing a research agenda through the power of collective intelligence. *Big Data and Society*, 1–21.

Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

Dennett, D. (1991). Real patterns. *Journal of Philosophy, 87*, 27–51.
Dennett, D. (1995). *Darwin's dangerous idea: evolution and the meanings of life*. New York: Simon & Schuster.
Diakopoulos, N. (2015). Algorithmic accountability: journalistic investigation of computational power structures. *Digital Journalism, 3*(3), 398–415.
Dutta, S. (2017) Do computers make better bank managers than humans? *The Conversation*.
Dworkin, R. (1977). *Taking rights seriously*. London: Duckworth.
Dworkin, R. (1986). *Law's empire*. London: Fontana Books.
Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*(1), 18–84.
Edwards, L. & Veale, M. (2018) Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy*.
Erdélyi, O.J. & Goldsmith, J. (2018) Regulating artificial intelligence: proposal for a global solution. *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*. Available at: http://www.aiesconference.com/wpcontent/papers/main/AIES_2018_paper_13.pdf.
Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St Martin's Press.
Fodor, J. A. (1981). Three cheers for propositional attitudes. In J. A. Fodor (Ed.), *RePresentations: philosophical essays on the foundations of cognitive science*. Cambridge: MIT Press.
Forssbæck, J., & Oxelheim, L. (2014). The multifaceted concept of transparency. In J. Forssbæck & L. Oxelheim (Eds.), *The Oxford handbook of economic and institutional transparency* (pp. 3–31). New York: Oxford University Press.
Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: speech acts* (pp. 41–58). New York: Academic Press.
Griffiths, J. (2016) New Zealand passport robot thinks this Asian man's eyes are closed. CNN.com December 9, 2016.
Hardt, M., Price, E. & Srebro, N. (2016) Equality of opportunity in supervised learning. *30th Conference on Neural Information Processing Systems* (NIPS 2016). Available at: https://arxiv.org/pdf/1610.02413v1.pdf.
Heald, D. (2006). Transparency as an instrumental value. In C. Hood & D. Heald (Eds.), *Transparency: the key to better governance?* (pp. 59–73). Oxford: Oxford University Press.
Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65–81.
Johnson, J.A. (2006). Technology and pragmatism: from value neutrality to value criticality. *SSRN Scholarly Paper, Rochester, NY: Social Science Research Network*. Available at: http://papers.ssrn.com/abstract=2154654.
Kleinberg, J., Mullainathan, S. & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *8th Conference on Innovations in Theoretical Computer Science* (ITCS 2017). Available at: https://arxiv.org/pdf/1609.05807.pdf.
Klingele, C. (2016). The promises and perils of evidence-based corrections. *Notre Dame Law Review, 91*(2), 537–584.
Langer, E., Blank, A. E., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: the role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology, 36*(6), 635–642.
Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016) How we analyzed the COMPAS recidivism algorithm. *ProPublica.org* May 23, 2016.
Leslie, S. (2017). The original sin of cognition: race, prejudice and generalization. *Journal of Philosophy, 114*(8), 393–421.
Levendowski, A. (2017) How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review* (forthcoming). Available at: https://ssrn.com/abstract=3024938.
Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6*, 539.
Lum, K. & Isaac, W. (2016) To predict and serve? Bias in police-recorded data. *Significance,* 14–19.
McEwen, R., Eldridge, J., & Caruso, D. (2018). Differential or deferential to media? The effect of prejudicial publicity on judge or jury. *International Journal of Evidence and Proof, 22*(2), 124–143.
Miller, T. (2017) Explanation in artificial intelligence: insights from the social sciences. Available at: https://arxiv.org/pdf/1706.07269.pdf.
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data and Society, 16*, 1–21.

Montavon, G., Bach, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with Deep Taylor decomposition. *Pattern Recognition, 65*, 211.

Muehlhauser (2013) Transparency in safety-critical systems. *Intelligence.org* August 15, 2013. Available at: https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/.

Nusser, S. (2009). *Robust learning in safety-related domains: machine learning methods for solving safety-related application problems*. Doctoral dissertation, Otto-von-Guericke-Universita¨t Magdeburg. Available at: https://pdfs.semanticscholar.org/48c2/e5641101a4e5250ad903828c02025d269a1a.pdf.

Oliver, W. M., & Batra, R. (2015). Standards of legitimacy in criminal negotiations. *Harvard Negotiation Law Review, 20*, 61–120.

Oswald, M. & Grace, J. (2016). Intelligence, policing and the use of algorithmic analysis: A freedom of information-based study. *Journal of Information Rights, Policy and Practice, 1*(1). Available at: https://journals.winchesteruniversitypress.org/index.php/jirpp/article/view/16.

Pasquale, F. (2014). *The black box society: the secret algorithms that control money and information*. Cambridge: Harvard University Press.

Piattelli-Palmarini, M. (1995). *La r'eforme du jugement ou comment ne plus se tromper*. Paris: Odile Jacob.

Plous, S. (2003a). The psychology of prejudice, stereotyping, and discrimination. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 3–48). New York: McGraw-Hill.

Plous, S. (2003b). *Understanding prejudice and discrimination*. New York: McGraw-Hill.

Pohl, J. (2008). Cognitive elements of human decision making Jens. In G. Phillips-Wren, N. Ichalkaranje, & L. C. Jain (Eds.), *Intelligent decision making: an AI-based approach* (pp. 3–40). Berlin: Springer.

Pomerol, J.-C., & Adam, F. (2008). Understanding human decision making: a fundamental step towards effective intelligent decision support. In G. Phillips-Wren, N. Ichalkaranje, & L. C. Jain (Eds.), *Intelligent decision making: an AI-based approach* (pp. 41–76). Berlin: Springer.

Prat, A. (2006). The more closely we are watched, the better we behave? In C. Hood & D. Heald (Eds.), *Transparency: the key to better governance?* (pp. 91–103). Oxford: Oxford University Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale: Lawrence Erlbaum Associates.

Schwab, K. (2016). *The fourth industrial revolution*. Geneva: Crown.

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues, 55*(4), 729–743.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge: MIT Press.

Tatman, R. (2016) Google's speech recognition has a gender bias. *Making Noise and Hearing Things*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124–1131.

Van Otterlo, M. (2013). A machine learning view on profiling. In M. Hildebrandt & K. de Vries (Eds.), *Privacy, due process and the computational turn: philosophers of law meet philosophers of technology* (pp. 41–64). Abingdon: Routledge.

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law and Security Review, 34*, 398–404.

Wachter, S., Mittelstadt, B. D., & Floridi, L. (2017a). Transparent, explainable, and accountable AI for robotics. *Science Robotics, 2*(6).

Wachter, S., Mittelstadt, B. D., & Floridi, L. (2017b). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76–99.

Waldron, J. (1990). *The law*. London: Routledge.